Information Theoretic Approaches to Rapid Discovery of Relationships in Large Climate Data Sets

Kevin H. Knuth[1], William B. Rossow[2]

1. NASA Ames Research Center, Moffett Field CA, 94035-1000
2. NASA Goddard Institute for Space Studies, New York NY, 10025

Mutual information as the asymptotic Bayesian measure of independence is an excellent starting point for investigating the existence of possible relationships among climate-relevant variables in large data sets. As mutual information is a nonlinear function of its arguments, it is not beholden to the assumption of a linear relationship between the variables in question and can reveal features missed in linear correlation analyses. However, as mutual information is symmetric in its arguments, it only has the ability to reveal the probability that two variables are related. It provides no information as to how they are related; specifically, causal interactions or a relation based on a common cause cannot be detected. For this reason we also investigate the utility of a related quantity called the transfer entropy. The transfer entropy can be written as a difference between mutual informations and has the capability to reveal whether and how the variables are causally related. The application of these information theoretic measures is tested on some familiar examples using data from the International Satellite Cloud Climatology Project (ISCCP) to identify relations between global cloud cover and other variables, including equatorial pacific sea surface temperature (SST), over seasonal and El Nino Southern Oscillation (ENSO) cycles.

Information Theoretic Approaches to Rapid Discovery of Relationships in Large Climate Data Sets

Kevin H. Knuth[1], William B. Rossow[2]

[1] NASA Ames Research Center, Computational Sciences Division, Moffett Field CA 94035 USA
[2] NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025 USA

B61A-0703

Abstract

Mutual information as the asymptotic Bayesian measure of independence is an excellent starting point for investigating the existence of possible relationships among climate-relevant variables in large data sets. As mutual information is a nonlinear function of its arguments, it is not beholden to the assumption of a linear relationship between the variables in question and can reveal features missed in linear correlation analyses. However, as mutual information is symmetric in its arguments, it only has the ability to reveal the probability that two variables are related. It provides no information as to how they are related; specifically, causal interactions or a relation based on a common cause cannot be detected. For this reason we also investigate the utility of a related quantity called the transfer entropy. The transfer entropy can be written as a difference between mutual informations and has the capability to reveal whether and how the variables are causally related. The application of these information theoretic measures is tested on some familiar examples using data from the International Satellite Cloud Climatology Project (ISCCP) to identify relations between global cloud cover and other variables, including equatorial pacific sea surface temperature (SST), over seasonal and El Nino Southern Oscillation (ENSO) cycles.

Introduction

In studying a complex dynamical system, one of the first approaches is to break it down into a set of dynamically coupled subsystems. In cases where this approach succeeds, we are able to develop a set of theories describing the behaviors of the subsystems. However, combining the theories of the parts to create a theory of the whole often represents an enormous challenge.

One attempt to create a theory of the whole is to describe the coupling among the subsystems. In linear systems this treatment is successful and its success in this regime continues to encourage application of the approach. However, in a nonlinear dynamical system it is often not even possible to define a coupling strength or to classify a given subsystem as driving or responding.

At this stage it is important to step back from the problem and reconsider our approach. Rather than focusing on driving systems and responding systems, we can examine how our state of knowledge about the state of the system at future times is affected by our knowledge about the present or past states of the various subsystems. This will provide much information regarding the interactions among the subsystems as well as giving us an idea as to the predictability of the various parts of the system.

To approach this we look at the mutual information between two subsystems. As this is the asymptotic Bayesian measure of independence [1], it will allow us to tease apart and identify dynamically independent subsystems. Last we examine transfer entropy [2, 3] as a measure of how information about one subsystem affects the state of knowledge about another subsystem at subsequent times.

Mutual Information

We can characterize the behavior of a system X by looking at the probability distribution over the set of states the system visits as it evolves in time. If a state is visited rarely, we might be surprised to find the system there. We can express the expectation (or lack of) to find the system in state x as

$$h(x) = \log \frac{1}{p(x)}$$

If we average this over all states, this gives us a measure of our expectation (or our uncertainty). This quantity is called the Shannon entropy [4]

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)}$$

If the system states can be described with multiple parameters, the entropy can still be computed by averaging over all possible states (here it is shown for states described by X and Y)

$$H(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{1}{p(x,y)}$$

Now, if we consider two subsystems X and Y, which together make up a larger system, we can compute what is called the **Mutual Information** (MI) by

$$MI(X,Y) = \left(H(X) + H(Y)\right) - H(X,Y)$$

Notice that this describes the difference between the uncertainty when they are treated separately and when they are treated jointly. If these two subsystems are independent of one another, the MI will be zero. However, if there is any interaction between these subsystems, the MI will be positive. This can perhaps be seen more clearly by writing it as

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)}$$

If X and Y are independent then the probability in the numerator factors and the MI is zero.

Transfer Entropy

However, as the MI is symmetric with respect to interchange of X and Y it cannot distinguish the direction of any causal influence as well as the effect of any common influence on both subsystems from a third system. A similar quantity called the **Transfer Entropy** can be written as a difference of MIs

$$T(X_{i+1} \mid X_i^{(k)}, Y_j^{(l)}) = MI(X_{i+1}, X_i^{(k)} \otimes Y_j^{(l)}) - MI(X_{i+1}, X_i^{(k)})$$

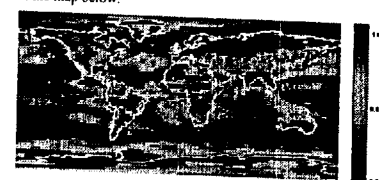
Where $X_i^{(k)}$ is the joint system defined by the set of states from X_{i-k+1} to X_i and $X_i^{(k)} \otimes Y_j^{(l)}$ is the system represented by the k previous states from X along with the l previous states from Y. Thus it measures the change in knowledge obtained from incorporating information about the subsystem Y. For $k = l = 1$ this can be nicely written in terms of Shannon entropies (although not as clearly interpretable)

$$T(X_{i+1} \mid X_i^{(k)}, Y_j^{(l)}) = -H(X_i) + H(X_i, Y_j) + H(X_i, X_{i+1}) - H(X_i, X_{i+1}, Y_j)$$

Seasonality and Cloud Cover

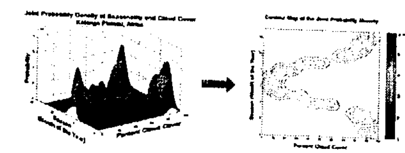
We used data (climate summary product C2) from the International Satellite Cloud Climatology Project (ISCCP) to investigate the ability of these information-theoretic techniques to efficiently and accurately discover relationships between seasonal change and global cloud cover.

The data consisted of monthly averages of percent cloud cover resulting in a time-series of 198 months of 6596 equal-area pixels each with side length of 280 km. The analysis was performed pixel-wise so that for each pixel: X = cloud cover percentages and Y = month of the year (seasonal state). The MI was computed for each pixel independently and is color-coded on the map below.

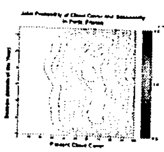


This method finds the Inter-Tropical Convection Zones, The Monsoon Regions, the Sea Ice off Antarctica, and cloud cover in the North Atlantic and Pacific. This figure can be directly compared to the PCA analysis performed by Rossow et al. [5].

Focusing on the high-MI region in the Katanga Plateau of southern Congo we see that the joint probability density is not factorable as indicated by the MI. The summer months are sunny and the winter months are cloudy.

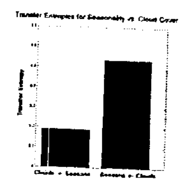


In contrast the cloud cover over Paris France has a low MI with respect to the seasonality. This is reflected by the highly factorable joint probability density to the right. Note that there is little dependence of cloud cover on seasonality.



The transfer entropies were found to be difficult to estimate with precision.

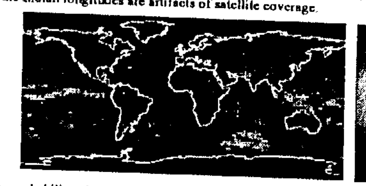
Using a Gaussian kernel density estimator, we found that the seasonality drove the cloud cover, which is reassuring as we know that the seasonality is an autonomous variable.



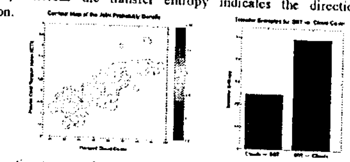
ENSO and Cloud Cover

The next example looks at the relationship between global cloud cover and Cold Tongue Index (CTI) [6], which describes the sea surface temperature anomalies in the eastern equatorial Pacific Ocean (6N-6S, 180-90W) and is indicative of ENSO variability [7]. Data is from [8].

The cloud cover affected by the SST variations lies mainly in the equatorial Pacific, along with an area in Indonesia. The highlighted areas in the Indian longitudes are artifacts of satellite coverage.



The probability density below shows how the two variables are co-dependent, whereas the transfer entropy indicates the direction of interaction.



Entropy estimates can be improved using Markov chain Monte Carlo. This approach will result in estimates with error bars. In addition, we will be able to directly test for independence of climate parameters.

References

1. Wolf, D.R. 1994. Mutual information as a Bayesian measure of independence. *LA-UR*, Vol. 94, 1994.
2. Schreiber, T. 2000. Measuring information transfer. *Phys Rev Lett* 85:461.
3. Kaiser A, Schreiber T. 2002. Information transfer in continuous processes. *Physica D* 166:43-62.
4. Shannon C.E., Weaver W. 1949. *The Mathematical Theory of Information*, University of Illinois Press, Urbana IL.
5. Rossow W.B., Walker A.W., Garder L.C. Comparison of ISCCP and other cloud amounts. J. Climate, 6:2394-2418.
6. Deser, C., Wallace J.M. 1990. Large-scale atmospheric circulation features of warm and cold episodes in the tropical Pacific. *J. Climate*, 3:1254-1281.
7. Deser, C., Wallace J.M. 1987. El Niño events and their relation to the Southern Oscillation: 1925-1986. *J. Geophys. Res.*, 92:14189-14196.
8. Mitchell, T. *http://tao.atmos.washington.edu/pacs/additional_analyses/sstanomtin6s18090w.html*

Funded by

Intelligent Data Understanding Project / Intelligent Systems Program
NASA Aerospace Technology Enterprise